THE EFFICIENCY OF BIAS–CORRECTED ESTIMATORS FOR NONPARAMETRIC KERNEL ESTIMATION BASED ON LOCAL ESTIMATING EQUATIONS

Göran Kauermann, Marlene Müller and Raymond J. Carroll *

June 2, 1997

Abstract

Stuetzle and Mittal (1979) for ordinary nonparametric kernel regression and Kauermann and Tutz (1996) for nonparametric generalized linear model kernel regression constructed estimators with lower order bias than the usual estimators, without the need for devices such as second derivative estimation and multiple bandwidths of different order. We derive a similar estimator in the context of local (multivariate) estimation based on estimating functions. As expected, this lower order bias is bought at a cost of increased variance. Surprisingly, when compared to ordinary kernel and local linear kernel estimators, the bias–corrected estimators increase variance by a factor independent of the problem, depending only on the kernel used. The variance increase is approximately 40% and more for kernels in standard use. However, the variance increase is still less than that incurred when undersmoothing a local quadratic regression estimator.

Key words and phrases: Bias Reduction; Bootstrap; Estimating Equations; Generalized Linear Models; Local Linear Regression; Nonparametric Regression.

Short title. Proportional Variance Inflation for Bias Reduction

*Göran Kauermann is Wissenschaftlicher Assistent, Fachgebiet Statistik und Wirtschaftsmathematik, Technischen Universität Berlin, Franklinstraße 28/29, D–10587 Berlin. Marlene Müller is Wissenschaftliche Assistentin, Institut für Statistik und Ökonometrie, Humboldt Universität zu Berlin, Spandauerstraße 1, D–10178 Berlin. Raymond J. Carroll is Professor of Statistics, Nutrition and Toxicology, Department of Statistics, Texas A&M University, College Station, TX 77843–3143. Carroll's research was supported by a grant from the National Cancer Institute (CA–57030), and was partially completed while visiting the Institut für Statistik und Ökonometrie, Sonderforschungsbereich 373, Humboldt Universität zu Berlin, with partial support from a senior Alexander von Humboldt Foundation research award.

1 INTRODUCTION

Nonparametric function estimation is greatly complicated by the problem of bias, and this has important consequences for inferential methods such as confidence bands and test statistics. For example, consider the problem of ordinary nonparametric regression to estimate the regression function $\Theta(x_0)$ of a response Y on a predictor X evaluated at a value x_0: $\Theta(x_0) = E(Y|X = x_0)$. Local linear regression estimators are solutions to the locally weighted estimating equation

$$0 = \sum_{i=1}^{n} w_i(x_0) \{Y_i - \beta_0 - \beta_1(X_i - x_0)\} (1, X_i - x_0)^T, \tag{1}$$

where the intercept is the regression estimate $\widehat{\Theta}(x_0, h)$ and the weights $w_i(x_0)$ increase with the distance of X_i to x_0. Consider the case that the weights are kernel weights $K_h(X_i - x_0) = h^{-1}K\{(X_i - x_0)/h\}$ with a symmetric kernel density function $K(\cdot)$. Let $\Theta^{(2)}(x)$ be the second derivative of the function $\Theta(x)$, let $f_x(x)$ be the density function of X and and suppose that the variance of Y given X is constant and equal to σ^2. Then it is well–known (Fan & Gijbels, 1996 is a convenient reference) that if x_0 is interior to the support of X, the bias and the variance are given approximately by

$$\text{bias}\left\{\widehat{\Theta}(x_0, h)\right\} \approx (1/2)h^2\Theta^{(2)}(x_0) \int z^2 K(z)dz; \tag{2}$$

$$\text{var}\left\{\widehat{\Theta}(x_0, h)\right\} \approx \sigma^2 \{nhf_x(x_0)\}^{-1} \int K^2(z)dz. \tag{3}$$

Results similar to (2)–(3) hold for ordinary kernel regression, generalized linear models (Fan, Heckman & Wand, 1995) and more generally for local estimating equations (Carroll, Ruppert & Welsh, 1998), see section 2 for definitions.

In practice, one must estimate the bandwidth, and this is usually done by minimizing an estimate of the mean squared error based on (2)–(3), see Ruppert (1998) for review. For these bandwidth estimators, for which h is proportional to $n^{-1/5}$, the squared bias and variance are set equal. The net effect is that while the estimators are optimal in a mean squared error sense, the (squared) bias is approximately the same as the variance.

Thus, while optimal bandwidth estimators give good function estimates, the fact that their squared bias approximately equals the variance has important implications for inferences. For example, if one ignores the bias, confidence intervals for the regression function at a point x_0 have coverage levels asymptotically smaller than the nominal. To overcome this problem, a standard technique is to estimate the second derivative function $\Theta^{(2)}(x_0)$, and then subtract the estimated bias from $\widehat{\Theta}_0(x_0, h)$, a technique employed either effectively or explicitly by Härdle & Bowman

(1988), Härdle & Marron (1991), Eubank & Speckman (1993) and Hall (1993), among others. The result is to remove the bias, and asymptotically not change the variance, because the h^2 in (2) is of small order, and any error in estimating the second derivative is of even lower asymptotic order.

There are a few difficulties with this general approach. First, all the above referenced papers include the need for a second bandwidth, which we will call g. The second bandwidth is needed effectively to estimate the second derivative function and hence necessarily converges to zero more slowly than the main bandwidth h. The second bandwidth g must be estimated as well, and this if usually much harder to accomplish than estimating h itself. Second, outside the context of estimating a mean function (Kauermann & Tutz, 1996; Carroll, Ruppert & Welsh, 1998), while it is possible to work out formulae similar to (2)–(3), it is neither clear how best to estimate the vector of second derivative functions (and their second bandwidths) nor whether such estimation leads to reasonable bias reduction properties in small samples.

To address these concerns, Kauermann & Tutz suggest an alternative method of bias reduction. While the method is given in its estimating equation form in section 2, here we derive it in ordinary kernel regression. The estimated regression function has the algebraic expression $\widehat{\Theta}(x_0, h) = \sum_1^n w_i(x_0) Y_i / \sum_1^n w_i(x_0)$, and from this one deduces that

$$\widehat{\Theta}(x_0, h) - \Theta(x_0) = b_n + E_n = \frac{\sum_{i=1}^n w_i(x_0) \{\Theta(X_i) - \Theta(x_0)\}}{\sum_{i=1}^n w_i(x_0)} + \frac{\sum_{i=1}^n w_i(x_0) \{Y_i - \Theta(X_i)\}}{\sum_{i=1}^n w_i(x_0)}.$$

The term b_n determines the bias, while the term E_n is a mean–zero random variable which determines the variance. The simplest device to estimate bias is to plug–in $\widehat{\Theta}(\cdot)$ for $\Theta(\cdot)$ in b_n, leading to the bias–corrected estimator

$$\begin{aligned} \widehat{\Theta}_c(x_0, h) &= \widehat{\Theta}(x_0, h) - \sum_{i=1}^n w_i(x_0) \left\{ \widehat{\Theta}(X_i) - \widehat{\Theta}(x_0) \right\} / \sum_{i=1}^n w_i(x_0) \qquad (4) \\ &= \sum_{i=1}^n w_i(x_0) \left\{ 2\widehat{\Theta}(x_0) - \widehat{\Theta}(X_i) \right\} / \sum_{i=1}^n w_i(x_0). \end{aligned}$$

The estimator (4) is a bias–corrected estimator which was called the twicing estimator by Stuetzle and Mittal (1979): it has bias of lower order than the usual kernel estimator. In section 2 we discuss the generalization of this estimator to the estimating equation context.

One would expect that bias–corrected estimators such as (4) should have larger variance than the ordinary estimator. We consider this question for estimating equations using local average and local linear kernel methods with symmetric kernel $K(\cdot)$. Our conclusion is surprising: *independent of the problem*, bias–corrected estimators are more variable by a factor depending only on the

kernel, namely

$$c(K) = \frac{\int K(z_2)K(z_3)\left\{2K(z_1) - K(z_2 - z_1)\right\}\left\{2K(z_1) - K(z_3 - z_1)dz_1 dz_2 dz_3\right\}}{\int K^2(z)dz}. \tag{5}$$

For the Gaussian kernel, $c(\text{Gaussian}) = 1.44$, while for the Epanechnikov kernel, $c(\text{Epanechnikov}) = 1.42$.

An alternative device to remove bias is direct undersmoothing, e.g., using a local quadratic regression with a bandwidth $h \sim n^{-1/5}$ from local linear regression. The increase in the variance, $d(K)$ say, from this can be computed from results of Ruppert & Wand (1994). For the Gaussian kernel, $d(\text{Gaussian}) = 1.69$, while for the Epanechnikov kernel, $d(\text{Epanechnikov}) = 2.08$. Both variance increases are larger than for the bias–corrected estimators (4).

2 LOCAL ESTIMATING EQUATIONS AND ESTIMATES OF BIAS

In parametric problems, estimation of a possibly vector–valued parameter Θ is typically based on an unbiased estimating function $\psi(\cdot)$, so that if the data are generically denoted by $\widetilde{\mathbf{Y}}_i$ $(i = 1, ..., n)$, then $\widehat{\Theta}$ is the solution to the estimating equation

$$0 = n^{-1}\sum_{i=1}^{n}\psi(\widetilde{\mathbf{Y}}_i, \widehat{\Theta}).$$

By an unbiased estimating function, we mean that $E\psi(\widetilde{\mathbf{Y}}, \Theta) = 0$. The choices of $\psi(\cdot)$ are well known. For example, when the data $\widetilde{\mathbf{Y}}$ consist only of a response Y and Θ is the mean of Y, $\psi(\widetilde{\mathbf{Y}}, \Theta) = Y - \Theta$. In generalized linear models, the data $\widetilde{\mathbf{Y}}$ consist of a response Y and covariates Z. The mean is $\mu(Z^T\Theta)$, the variance is proportional to $V(Z^T\Theta)$, and the estimating function is the quasilikelihood score $\psi(\widetilde{\mathbf{Y}}, \Theta) = Z\mu^{(1)}(Z^T\Theta)\left\{Y - \mu(Z^T\Theta)\right\}/V(Z^T\Theta)$, where $\mu^{(1)}(\cdot)$ is the first derivative of the function $\mu(\cdot)$.

Nonparametric regression can be thought of as a varying coefficient model (Kauermann & Tutz, 1996), where the coefficient Θ varies with a covariate X. An estimate of $\Theta(x_0)$ can be obtained using local polynomials of order $p \geq 0$ as follows. Define $G_p(v) = (1, v, ..., v^p)^T$, and let the weights $w_i(x_0)$ be as in Section 1. Suppose that $\Theta(x)$ is a vector of length q. Let $\otimes$ is the Kronecker product, so that for example $(a, c) \otimes (b1, b2, b3) = (ab_1, ab_2, ab_3, cb_1, cb_2, cb_3)$. Define $\mathcal{B}^T = (\beta_0^T, ..., \beta_p^T)$. Then $\widehat{\Theta}(x_0)$ is the intercept $\widehat{\beta}_0$ in the solution to the equation

$$0 = n^{-1}\sum_{i=1}^{n} w_i(x_0)G_p(X_i - x_0) \otimes \psi\left[\widetilde{\mathbf{Y}}_i, \left\{G_p^T(X_i - x_0) \otimes I_q\right\}\widehat{\mathcal{B}}\right]. \tag{6}$$

For these local polynomial estimators, formulae similar to (2)–(3) hold. In fact, if $p = 1$, the bias is still given by (2), while the variance is the same as (3) except that σ^2 is replaced by $\left\{g^{-1}(x)\ell(x)g^{-T}(x)\right\}$, where $g^{-T}(\cdot)$ is the transpose of $g^{-1}(\cdot)$, $\chi(\widetilde{\mathbf{Y}}, \Theta) = -(\partial/\partial\Theta)\psi(\widetilde{\mathbf{Y}}, \Theta)$, $g(x) = E[\chi\{\widetilde{\mathbf{Y}}, \Theta(X)\}|X = x]$ and $\ell(x) = E[\psi\{\widetilde{\mathbf{Y}}, \Theta(X)\}\psi^T\{\widetilde{\mathbf{Y}}, \Theta(X)\}|X = x]$.

A bias–corrected estimator is constructed as follows. Define

$$B_n(x_0) = n^{-1}\sum_{i=1}^{n} w_i(x_0)\left[G_p(X_i - x_0)G_p^T(X_i - x_0) \otimes \chi\left\{\widetilde{\mathbf{Y}}_i, \Theta(X_i)\right\}\right].$$

Then, by a Taylor series expansion,

$$\widehat{\Theta}(x_0) - \Theta(x_0) \approx b_n + E_n, \tag{7}$$

where

$$\begin{aligned} E_n &= e_p B_n^{-1}(x_0) n^{-1}\sum_{i=1}^{n} w_i(x_0) G_p(X_i - x_0) \otimes \psi\left\{\widetilde{\mathbf{Y}}_i, \Theta(X_i)\right\} \\ b_n &= e_p B_n^{-1}(x_0) n^{-1}\sum_{i=1}^{n} w_i(x_0) G_p(X_i - x_0) \otimes \left(\psi\left[\widetilde{\mathbf{Y}}_i, \left\{G_p^T(X_i - x_0) \otimes I_q\right\}\mathcal{B}\right] - \psi\left\{\widetilde{\mathbf{Y}}_i, \Theta(X_i)\right\}\right). \end{aligned}$$

with e_p be the $q \times q(p+1)$ matrix of zeros except that the first $q \times q$ submatrix is the identity matrix. Just as in (4), the idea is to estimate the terms in the bias, leading to the estimator

$$\begin{aligned} \widehat{\Theta}_c(x_0) &= \widehat{\Theta}(x_0) - e_p\widehat{B}_n^{-1}(x_0)\mathcal{C}_n(x_0); \qquad (8) \\ \mathcal{C}_n(x_0) &= n^{-1}\sum_{i=1}^{n} w_i(x_0) G_p(X_i - x_0) \otimes \left(\psi\left[\widetilde{\mathbf{Y}}_i, \left\{G_p^T(X_i - x_0) \otimes I_q\right\}\widehat{\mathcal{B}}\right] - \psi\left\{\widetilde{\mathbf{Y}}_i, \widehat{\Theta}(X_i)\right\}\right); \\ \widehat{B}_n(x_0) &= n^{-1}\sum_{i=1}^{n} w_i(x_0)\left[G_p(X_i - x_0)G_p^T(X_i - x_0) \otimes \chi\left\{\widetilde{\mathbf{Y}}_i, \widehat{\Theta}(X_i)\right\}\right]. \end{aligned}$$

In the case of the likelihood score $\psi(\cdot)$ with local averages, $p = 0$, $G_p(v) = 1$ and the bias–corrected estimator derived from (8) reproduces the bias corrected estimator of Kauermann & Tutz.

In the appendix, we show that for local averages ($p = 0$) and local linear smoothing ($p = 1$), with bandwidth $h \sim n^{-1/5}$ the bias corrected estimator is always more variable asymptotically than the uncorrected estimator (6) by the factor (5).

3 DISCUSSION

There are two general ways to correct for bias in nonparametric regression: (a) estimate the second derivative function directly and subtract a multiple of it from the usual estimator; and (b) bias–correct indirectly either by undersmoothing (applying a local–linear bandwidth to a local quadratic

estimator) or the twicing technique. The major difficulty with method (a) is the need for a second bandwidth. We have shown that methods (b) are more variable than method (a), by a constant factor independent of the problem. Between the two possibilities in method (b), the twicing estimator is asymptotically less variable.

The twicing estimator shows another advantage with respect to application. If the bias corrected estimators (4) and (8) are used, it is simple to estimate their variance: take any variance estimator for an uncorrected regression function which is already typically available in the literature, and multiply it by the variance inflation factor (5).

REFERENCES

Carroll, R. J., Ruppert, D. & Welsh, A. (1998). Nonparametric estimation via local estimating equations, with applications to nutrition calibration. *Journal of the American Statistical Association*, to appear.

Eubank, R. L. & Speckman, P. L. (1993). Confidence bands in nonparametric regression. *Journal of the American Statistical Association*, 88, 1287–1301.

Fan, J., Heckman, N. E. & Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models. *Journal of the American Statistical Association*, 90, 141–150.

Fan, J. & Gijbels, I. (1996): *Local Polynomial Modeling and its Applications.* London: Chapman & Hall.

Hall, P. (1993). On Edgeworth expansion and bootstrap confidence bands in nonparametric curve estimation. *Journal of the Royal Statistical Society, Series B*, 55, 291–304.

Härdle, W. & Bowman, A. W. (1988). Bootstrapping nonparametric regression: local adaptive smoothing and confidence bands. *Journal of the American Statistical Association*, 83, 102–110.

Härdle, W. & Marron, J. S. (1991). Bootstrap simultaneous error bars for nonparametric regression. *Annals of Statistics*, 19, 778–796.

Kauermann, G. & Tutz, G. (1996). On model diagnostics and bootstrapping in varying coefficient models. Submitted.

Ruppert, D. (1998). Empirical bias bandwidth selection. *Journal of the American Statistical Association*, to appear.

Ruppert, D., & Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Annals of Statistics*, 22, 1346–1370.

Stuetzle, W. & Mittal, Y. (1979). Some comments on the asymptotic behavior of robust smoothers. In *Smoothing Techniques for Curve Estimation*, editors T. Gasser and M. Rosenblatt, Springer Lecture Notes, 757, 191–195.

Appendix A PROOFS OF THEOREMS

In what follows, we will assume that $h \sim n^{-1/5}$, and we will use the notation $\approx$ to mean equality to terms of order $o_p(h^2)$. Recall that $f_x(\cdot)$ is the density of X. The kernel function $K(\cdot)$ is symmetric. Define $g(x) = E[\chi\{\widetilde{\mathbf{Y}}, \Theta(X)\}|X = x]$ and $\ell(x) = E[\psi\{\widetilde{\mathbf{Y}}, \Theta(X)\}\psi^T\{\widetilde{\mathbf{Y}}, \Theta(X)\}|X = x]$.

Our argument here is heuristic but can be justified under strong regularity conditions, e.g. X and K are compactly supported, the density of X is bounded away from zero on its support, etc. We provide details in the case of local averages, i.e., $p = 0$ in (6). The arguments are similar in the local linear case ($p = 1$) because the expansion (9) given below still holds in this case, but with (10) replaced by (2).

Carroll, et al. (1998) show that

$$\begin{aligned}
\widehat{\Theta}(x) - \Theta(x) &\approx C_n(x) \approx h^2 r(x) + D_n(x), \text{ where} \qquad (9)\\
B_n(x) &= n^{-1}\sum_{i=1}^n K_h(X_i - x)\chi\{\widetilde{\mathbf{Y}}_i, \Theta(x)\};\\
C_n(x) &= B_n^{-1}(x)n^{-1}\sum_{i=1}^n K_h(X_i - x)\psi\{\widetilde{\mathbf{Y}}_i, \Theta(x)\};\\
D_n(x) &= B_n^{-1}(x)n^{-1}\sum_{i=1}^n K_h(X_i - x)\psi\{\widetilde{\mathbf{Y}}_i, \Theta(X_i)\};\\
r(x) &= \{(1/2)\Theta^{(2)}(x) + f_x^{(1)}(x)\Theta^{(1)}(x)/f_x(x)\}\int z^2K(z)dz. \qquad (10)
\end{aligned}$$

The variance of $\widehat{\Theta}(x)$ is approximately

$$\Sigma = \{nhf_x(x)\}^{-1}\int K^2(z)dz\left\{g^{-1}(x)\ell(x)g^{-T}(x)\right\},$$

where $g^{-T}(\cdot)$ is the transpose of $g^{-1}(\cdot)$.

Using these expansions, we have that

$$\begin{aligned}
&\widehat{\Theta}_c(x_0) - \Theta(x_0)\\
&\approx B_n^{-1}(x_0)n^{-1}\sum_{i=1}^n K_h(X_i - x_0)\left[\psi\{\widetilde{\mathbf{Y}}_i, \widehat{\Theta}(X_i)\} - \psi\{\widetilde{\mathbf{Y}}_i, \widehat{\Theta}(x_0)\} + \chi\{\widetilde{\mathbf{Y}}_i, \Theta(x_0)\}\{\widehat{\Theta}(x_0) - \Theta(x_0)\}\right]\\
&\approx B_n^{-1}(x_0)n^{-1}\sum_{i=1}^n K_h(X_i - x_0)\chi\{\widetilde{\mathbf{Y}}_i, \Theta(x_0)\}[2\{\widehat{\Theta}(x_0) - \Theta(x_0)\} - \{\widehat{\Theta}(X_i) - \Theta(x_0)\}]\\
&\quad + B_n^{-1}(x_0)n^{-1}\sum_{i=1}^n K_h(X_i - x_0)[\psi\{\widetilde{\mathbf{Y}}_i, \Theta(X_i)\} - \psi\{\widetilde{\mathbf{Y}}_i, \Theta(x_0)\}]\\
&\approx B_n^{-1}(x_0)n^{-1}\sum_{i=1}^n K_h(X_i - x_0)\chi\{\widetilde{\mathbf{Y}}_i, \Theta(x_0)\}\{2D_n(x_0) - D_n(X_i)\}\\
&\quad + h^2B_n^{-1}(x_0)n^{-1}\sum_{i=1}^n K_h(X_i - x_0)\chi\{\widetilde{\mathbf{Y}}_i, \Theta(x_0)\}\{2r(x_0) - r(X_i)\}
\end{aligned}$$

$$+B_n^{-1}(x_0)n^{-1}\sum_{i=1}^{n}K_h(X_i-x_0)\{\psi\{\widetilde{\mathbf{Y}}_i,\Theta(X_i)\}-\psi\{\widetilde{\mathbf{Y}}_i,\Theta(x_0)\}\}$$
$$=G_{n1}+G_{n2}+G_{n3}.$$

It is easily shown that $G_{n2}\approx h^2r(x_0)$. Further, by calculations of first and second moments, it follows that $G_{n3}\approx -h^2r(x_0)$. Finally, we have that $B_n(x)=f_x(x)g(x)+O_p(h^2)$. From this, it follows that

$$\widehat{\Theta}_c(x_0)-\Theta(x_0)\approx\{f_x(x_0)g(x_0)\}^{-1}n^{-1}\sum_{i=1}^{n}K_h(X_i-x_0)\chi\{\widetilde{\mathbf{Y}}_i,\Theta(x_0)\}\{2D_n(x_0)-D_n(X_i)\}.$$

Now define

$$L_{ni}(x)=\{f_x(x)g(x)\}^{-1}n^{-1}\sum_{j=1}^{n}K_h(X_j-x)\psi\{\widetilde{\mathbf{Y}}_j,\Theta(X_j)\}$$

Then it is easily seen that

$$\widehat{\Theta}_c(x_0)-\Theta(x_0)\approx\{f_x(x_0)g(x_0)\}^{-1}n^{-1}\sum_{i=1}^{n}K_h(X_i-x_0)\chi\{\widetilde{\mathbf{Y}}_i,\Theta(x_0)\}\{2L_{ni}(x_0)-L_{ni}(X_i)\}. \quad (11)$$

Now write out (11) into a double sum in i and j, interchange the indices, eliminate the terms in which i and j are equal since they are of order $(nh)^{-1}=o_p(h^2)$ and use the assumption that $K(\cdot)$ is symmetric to get that

$$\widehat{\Theta}_c(x_0)-\Theta(x_0) \approx \{f_x(x_0)g(x_0)\}^{-1}n^{-1}\sum_{i=1}^{n}\psi\{\widetilde{\mathbf{Y}}_i,\Theta(X_i)\}M_n(x_0,X_i), \text{ where} \quad (12)$$
$$M_n(x_0,X_i) = n^{-1}\sum_{j=1,j\neq i}^{n}K_h(X_j-x_0)\chi\{\widetilde{\mathbf{Y}}_j,\Theta(x_0)\}\left\{\frac{2K_h(X_i-x_0)}{f_x(x_0)g(x_0)}-\frac{K_h(X_j-X_i)}{f_x(X_j)g(X_j)}\right\}$$

Recall the definition of $c(K)$ in (5). The right side of (12) is a mean zero random variable, and its variance is easily calculated to be $c(K)\Sigma\{1+o(1)\}$, as claimed.